Abstracts Embeddings Evaluation

A Case Study of Artificial Intelligence and Medical Imaging for the COVID-19 Infection

Giovanni Zurlo, Elisabetta Ronchieri

September 18, 2023

ML_INFN, Bologna, IT

Table of Contents

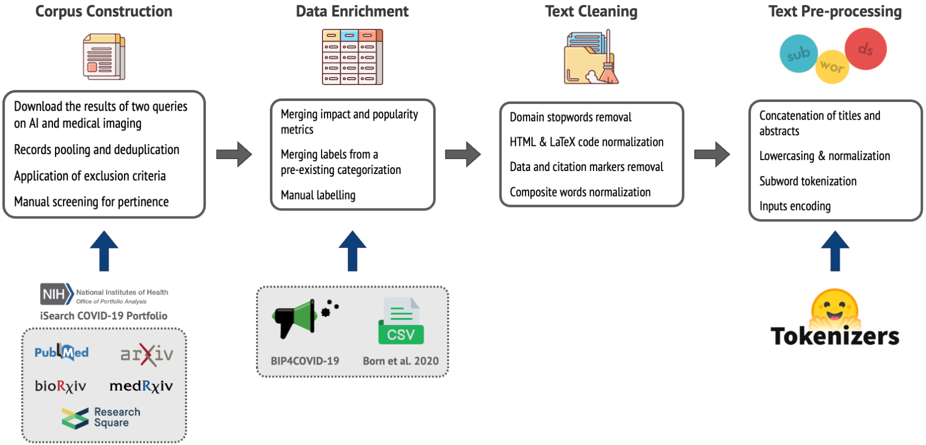
Problem Domain

1 Introduction

- The SARS-CoV-2 pandemic triggered unprecedented research efforts across various disciplines.
- This study delves into the collaborative prospects of artificial intelligence (AI) and medical imaging to expedite the analysis of scientific COVID-19 articles on larger scale.
- By harnessing the capabilities of natural language processing (NLP) and contextualized vector representations, the investigation scrutinizes the potential of popular biomedical transformer-based models to capture the semantic attributes in the medical imaging literature.

Table of Contents

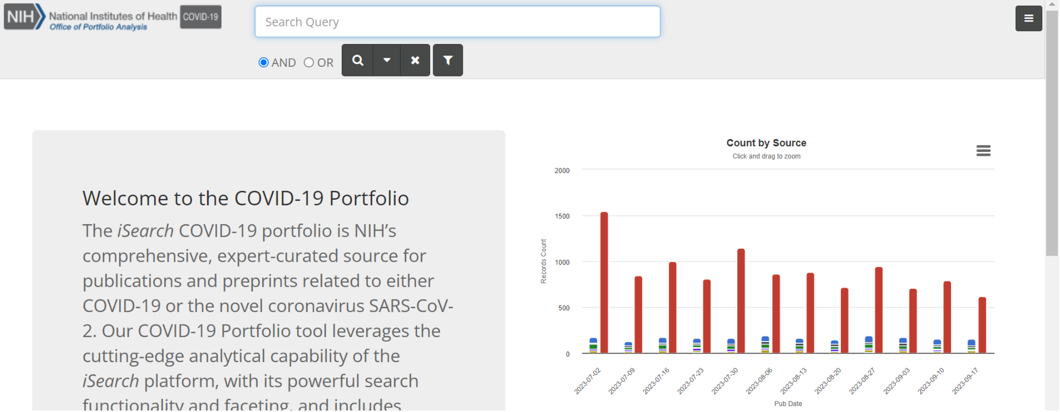
Data Collection Workflow



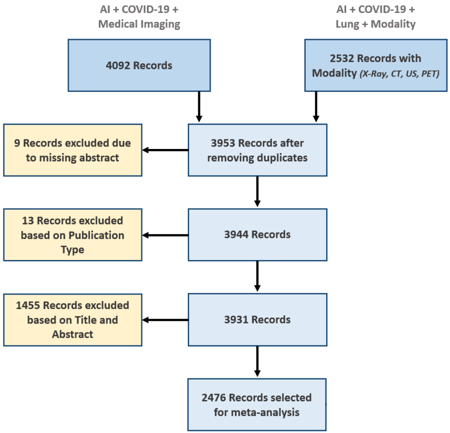
Papers Sources

2 Methodology

Considered the National Institutes of Health (NIH)'s iSearch COVID-19 Portfolio



PRISMA-based Corpus Definition



1. Broad query

AI AND COVID-19 AND 'Medical Imaging'

2. Modality-specific query

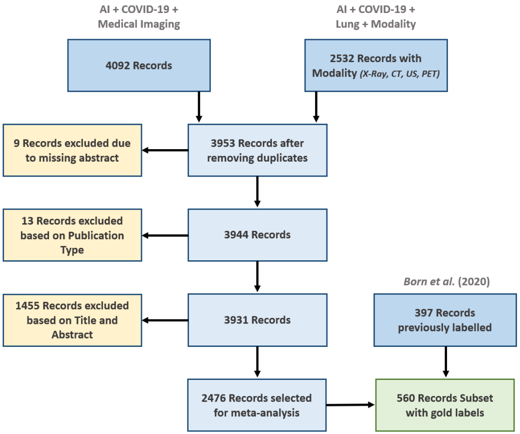
AI AND COVID-19 AND Lung AND (CT OR CXR OR US OR PET)

- CT Computerized Tomography
- CXR Chest X-Ray
- US Ultrasound
- PET Positron Emission Tomography

Collected papers $\in$ period Jan 1, 2020 - May 27, 2023

Used Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) to select papers

Data Enrichment



Identified 560 gold-papers

- 163 papers manually labelized to address the issue of class imbalance
- 397 papers derived by Born et al. (2020) (see ref [6] in the paper) based on a supplementary dataset titled 'Detailed results of systematic meta-analysis', merged by using title and already labelized

Data Enrichment - Labeling Assignment

Chose to adopt the tasks and modalities classification framework already adopted in Born et al. (2020) (see ref [6] in the paper)

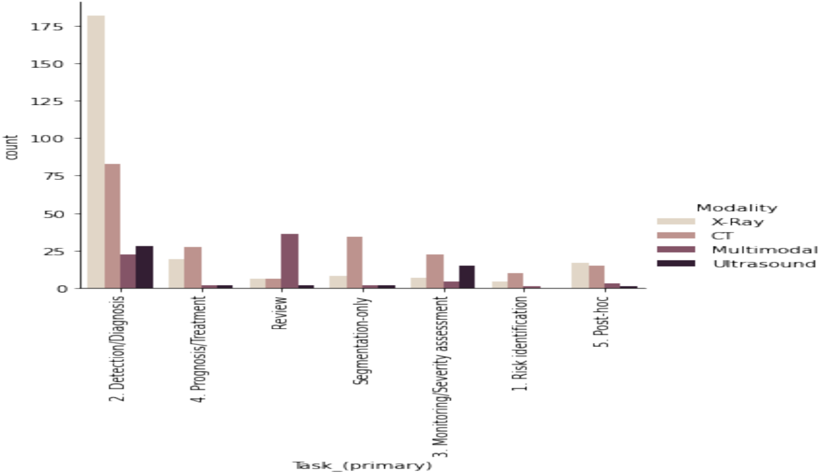
Primary Task

- Detection/Diagnosis
- Monitoring/Severity
- Assessment
- Post-Hoc
- Prognosis/Treatment
- Review
- Risk Identification
- Segmentation-only (for lung tissue or other disease features without any clinically relevant downstream tasks)

Imaging Modality

- CT
- CXR
- Lung US
- Multimodal

Data Enrichment - Labeling Assignment



15 Models

12 Bidirectional Encoder Representation (BERT) models

- original BERT in its base and large versions
- SciBERT
- BioBERT in its base and large versions
- PubMedBERT in its base and large versions
- CORD-19 BERT
- COVID SciBERT
- ClinicalCovidBERT
- RadBERT
- BioCovidBERT

3 SPECTER models

- standard model
- two-others with task-specific adapters

Models Distinctions Summary

Model	Training Corpus	Weights Initialization	Vocabulary	Details on Training Corpus
BERT$_{base}$	Wiki+Books	From-scratch	Derived from corpus	800M + 2.5B words, 1M steps
SciBERT	SemanticScholar Full-Texts	BERT$_{base}$	Derived from corpus	1.14M Full-Texts, 18% from computer science and 82% from broad biomedical domain
BioBERT$_{base}$	PubMed abstracts	BERT$_{Base}$	Same as BERT$_{Base}$	Updated 2019. 4.5B Words, 1M steps
PubMedBERT$_{base}$	PubMed abstracts + PMC Full-Texts	From-scratch	Derived from corpus	Updated Feb. 2020. 16.8B Words, 100K steps
CORD-19 BERT	CORD-19 dataset	BERT$_{Base}$	Same as BERT$_{Base}$	Updated Early 2020
CovidSciBERT	CORD-19 dataset	SciBERT	Extended from SciBERT	Updated Early 2020
ClinicalCovidBERT	CORD-19 dataset	Bio+Clinical BERT [1]	Same as BERT$_{Base}$	Full-Texts updated June 2020, 150K steps
RadBERT	Radiology Reports	BioBERT$_{base}$	Same as BERT$_{Base}$	4M reports from 600K unique patients treated at Stanford Health Care from 1992 to 2014
SPECTER 2	6M Triplets of Papers Citations	SciBERT	Same as SciBERT	Extended version of the `cite_prediction` dataset from [31]
BERT$_{large}$	Wiki+Books	From-scratch	Derived from corpus	800M + 2.5B words, 1M steps
BioBERT$_{large}$	PubMed abstracts	BERT$_{Large}$	Derived from corpus	Updated 2019. 4.5B Words, 1M steps
PubMedBERT$_{large}$	PubMed abstracts	From-scratch	Derived from corpus	Updated Feb. 2020. 3.2B Words, 100K steps
BioCovidBERT	CORD-19 dataset	BioBERT$_{large}$	Same as BioBERT$_{large}$	Full-Texts updated June 2020, 200K steps

BERTbase example

- Each title + abstract pair gets concatenated.
- Any BERT$_{base}$ model encodes each pair in 768-dimensional latent space.
- The first token is a special classification token [CLS].
- The separator token [SEP] marks its end and separates titles from abstracts.
- The context window is fixed at 512 tokens (almost 300-400 words), causing a truncation for longer inputs.
  - Our dataset adheres this constraint of the context window.
  - Few records required truncation.

Extraction Strategies

- Our goal is to obtain a singular vector representation for each Text + Abstract, no one for each token.
- Three extraction strategies considered:
  - the first two involve extracting the final hidden state representation of the [CLS] token and the trailing [SEP] token
  - the third uses the mean-pooling strategy based on the second-to-last hidden states.

Performance Metrics

- Accuracy is computed for a k-Nearest Neighbors (kNN) classifier to provide an evaluation of embedding quality.
- All kNN-based metrics involved k=6 or k=13 exact nearest neighbors.
  - ’KneihborsClassifier’ class from Scikit-Learn 1.2.2
  - All parameters were chosen performing a cross-validated grid search:
    - algorithm=’auto’, weights=’distance’, distance=’cosine’
- To predict each test paper’s label, kNN takes a weighted majority vote among the paper’s NNs’labels in the training set.
- Neibhbors are weighted by the inverse of their cosine distance.

Accuracy Details

- For the accuracy, cross-validated values were averaged over the same 10-fold split.
- Additionally, a balanced version of accuracy was computed.
- The chance-level accuracy was calculated by using 'DummyClassifier' with strategy='stratified' to ignore the input features.

Table of Contents

Quality metrics for the embeddings (Imaging Modality Prediction)

3 Results

10-fold kNN classification accuracy and balanced accuracy.
Hyperparameters: $k = 13$, $weights = distance$, $distance = cosine$.

Model	Accuracy (%)			Balanced Accuracy (%)		
	[CLS]	[SEP]	AVG	[CLS]	[SEP]	AVG
BERT_{base}	54.3	57.7	61.3	38.6	43.3	49.9
SciBERT	58.2	56.4	63.6	43.9	41	48.2
BioBERT_{base}	53.2	65.2	61.1	36	52.6	46.2
PubMedBERT_{base}	57.7	74.5	64.3	42.9	58.6	50.1
CORD-19 BERT	56.4	*52.9*	60.2	43.0	*37.6*	44.9
CovidSciBERT	64.5	60.5	62.7	49.9	47.9	50.6
ClinicalCovidBERT	65.4	64.5	63.4	53.2	50.8	49.6
RadBERT	57.9	57.9	*58.2*	38	38	*39.7*
SPECTER 2	**82.5**	**83.8**	68.8	**75.3**	**76.7**	57.8
BERT_{large}	*50*	58.8	60.2	*34.7*	43.4	44.8
BioBERT_{large}	54.4	62.1	65.5	39.8	47.2	51.3
$\text{PubMedBERT}_{large}$	57	61.3	60.4	40.3	44.2	45.3
BioCovidBERT	69.5	64.8	**69.6**	53.8	49	**58.3**
Chance Level		35.5 ±13			24.8 ±9	

SPECTER employs the [CLS] token, but we also applied the others for consistency.
$\text{BioCovidBERT}_{large}$ slightly outperformed with AVG pooling strategy due to its continual pre-trained on a COVID-19-based corpus.

Quality metrics for the embeddings (Task Prediction)

3 Results

10-fold kNN classification accuracy and balanced accuracy.
Hyperparameters: $k = 6$, $weights = distance$, $distance = cosine$.

Model	Accuracy (%)			Balanced Accuracy (%)		
	[CLS]	[SEP]	AVG	[CLS]	[SEP]	AVG
BERT$_{base}$	*59.6*	*58.2*	64.8	27	28.4	33.9
SciBERT	62.7	63	68.6	33.3	31.5	38.3
BioBERT$_{base}$	63.9	70.2	69.6	28.6	40.5	40.2
PubMedBERT$_{base}$	66.8	70.7	67.5	34.7	42.3	36.9
CORD-19 BERT	65.0	60.7	65.9	33.7	*25.9*	34.4
CovidSciBERT	70.2	70.2	71.8	42.3	42.4	45.1
ClinicalCovid BERT	70.9	71.3	70	43.6	46.7	40.9
RadBERT	60.9	60.9	*61.2*	26.5	26.5	*26.4*
SPECTER 2	**75.4**	**74.5**	**74.1**	**56.6**	**55.9**	**51.5**
BERT$_{large}$	60	64.1	66.6	*26.2*	34	38.5
BioBERT$_{large}$	62	68.6	67.3	28.7	37.7	36
PubMedBERT$_{large}$	63	67.7	68.9	30	36.1	38.4
BioCovidBERT	66.6	68.2	70.9	37	39.5	42.5
Chance Level		36.1 ±6			14.8 ±9	

The Balanced accuracy scores decreased due to the presence of stronger class imbalance and lower recall values for ’post-hoc’ and ’risk identification’ classes.

Table of Contents

Conclusions

- A first version of a medical imaging dataset for the COVID-19 infection has been defined by following the PRISMA procedure in order to evaluate embeddings quality for abstracts texts.
  - Extrinsic evaluation fails if the embeddings are trained to serve in a wide range of different tasks.
- We have labelized entries according to the primary task and the imaging modality.
- The SPECTER model emerges as the best model with respect to accuracy and balanced accuracy in task prediction diverse across diverse extraction strategies.
- Data and code of the paper are available at `https://github.com/zurlog/abs-embeddings-eval`.

Future Work

- To improve the annotation process of our original dataset
  - using a combination of automated tools and manual assessment.
- To collect more labeled entries in order to improve the training set sample size.
- To keep it updated.

Abstracts Embeddings Evaluation

A Case Study of Artificial Intelligence and Medical Imaging for the COVID-19 Infection

Thank you for listening!

Any questions?